Apprentissage par Renforcement et Réduction de la dimensionnalité : une Approche Bio-inspirée

Frédéric Alexandre, Nishal Shah

Apprentissage par Renforcement et Réduction de la dimensionnalité : une Approche Bio-inspirée

Frederic ALEXANDRE[1] et Nishal Shah[2]

[1] Centre INRIA de Nancy,
frederic.alexandre@inria.fr
http://www.loria.fr/~falex
[2] LORIA, Nancy Universités

Résumé : Depuis les travaux fondateurs de Schultz (Schultz et al., 1997), une structure cérébrale appelée Ganglions de la Base est reconnue comme possible substrat neuronal de l'Apprentissage par Renforcement. En conséquence, de nombreux travaux ont été menés pour préciser cette analogie, tant au niveau fonctionnel qu'anatomique. Un élément architectural notable a cependant été très peu pris en compte : la formidable réduction de dimensionnalité qui s'opère entre l'entrée et la sortie des Ganglions de la Base. Seul un auteur (Bar-Gad et al., 2003) a proposé que cette transformation pourrait être équivalente à une analyse en composantes principales, mais l'analyse des conséquences de cette hypothèse est restée superficielle. C'est dans ce cadre que nous avons choisi de reprendre ces travaux et de les étendre à un cadre plus fonctionnel et plus bio-inspiré afin d'étudier si cette hypothèse reste valide dans ce contexte plus réaliste. Nous montrons en particulier tout l'intérêt qu'il y a à ne pas simplement regarder cette transformation mais aussi à fermer la boucle, c'est à dire à l'utiliser pour la sélection de l'action.

Mots-clés : Apprentissage par Renforcement, Réduction de la Dimensionnalité, Modélisation bio-inspirée, Neurosciences Computationnelles.

1. Biologie de l'Apprentissage par Renforcement

1.1. Tableau général

Dans le domaine de la modélisation de fonctions adaptatives inspirée du fonctionnement cérébral, les Neurosciences Computationnelles se sont beaucoup intéressées aux propriétés du cortex (Fix et al., 2007). En résumé, la partie postérieure au sillon central du cortex représente son pôle sensoriel et est caractérisée par ses capacités d'auto-organisation de l'information sensorielle, par construction de cartes topologiques par apprentissage non-supervisé. Le modèle Self-Organizing-Maps (SOM) de Kohonen (Kohonen, 1997) et ses relations possibles à des modèles d'apprentissage statistique

comme les K-means en sont une parfaite illustration. La partie antérieure au sillon central du cortex (aussi appelée cortex frontal) représente son pôle moteur et est étudiée pour la modélisation de l'activité motrice, par exemple en robotique autonome (Frezza et Alexandre, 2002) et plus généralement pour l'organisation temporelle du comportement. Enfin, de nombreuses tâches sensori-motrices ont été modélisées par l'association de ces deux pôles (voir par exemple (Fix, 2007) pour le cas visuomoteur).

Le cortex est également un objet d'attention important de la part des modélisateurs car il représente l'une des plus grosses structures chez le primate (10^9 neurones) et qu'il est décrit en neuropsychologie comme le siège des fonctions cognitives les plus évoluées. Pour ce qui concerne l'apprentissage par renforcement (Sutton et Barto, 1998) et le problème de la sélection de l'action (en résumé, étant donnée la perception courante, sélectionner l'action qui maximise l'espérance de récompense), qui compte assurément parmi les fonctions cognitives les plus avancées, on pourrait penser que le cortex a tous les éléments pour prendre en charge ce problème, d'autant que les cortex postérieur et frontal incluent des aires spécifiques à la représentation interoceptive du corps et donc à la représentation de la récompense (Craig, 2009). Cependant, le cortex est également caractérisé par ses motifs de calcul locaux (chaque neurone n'est lié qu'à 10^3 ou 10^4 neurones corticaux, parmi les 10^9 cibles potentielles) et privilégie en interne les apprentissages sensorimoteurs stables (Goodale et Humphrey, 1998), à la différence du caractère hautement dynamique et changeant de l'apprentissage par renforcement (Sutton et Barto, 1998), même si certaines régions du cortex frontal sont également caractérisées par la volatilité de leurs représentations dans la planification de l'action (Cisek, 2005). Peter Redgrave et ses collègues proposent de résoudre ce dilemme (Redgrave et al., 1999) en postulant que les ganglions de la base (Basal Ganglia en anglais : BG), groupe de noyaux sous-corticaux interconnectés, fournissent, dans une boucle qu'ils font avec le cortex et le thalamus, le substrat physiologique à même de s'associer avec le cortex pour les tâches de sélection de l'action, en réalisant en particulier l'apprentissage par renforcement.

1.2. Les ganglions de la base

Les ganglions de la base (BG) sont en effet présentés par (Redgrave et al., 1999) comme un "interrupteur adaptatif" permettant une sélection de l'action motivée par une évaluation de la récompense prédite. En particulier, les deux boucles qui les composent permettent une analogie directe avec l'architecture d'acteur-critique (Joel et al., 2002), un des algorithmes fondamentaux de l'apprentissage par renforcement où l'acteur sélectionne la meilleure action à partir des perceptions courantes et des connaissances actuelles et le critique prédit la récompense espérée à partir des mêmes éléments et dont les erreurs de prédiction permettent de corriger les deux agents (Sutton et Barto, 1998).

La boucle basale (cortex-BG-thalamus-cortex) correspond à l'acteur. Cette boucle principale reçoit des entrées de la majeure partie du cortex (postérieur et frontal) dans sa structure d'entrée principale, le Striatum, grosse structure composée chez les primates de 10^7 neurones. Il y a également d'autres structures d'entrée comme le noyau sous-thalamique (STN), de plus petite taille, que nous ne décrirons pas ici. La couche de sortie est composée de deux structures que nous ne différencierons pas : GPi/SNr. Par défaut, ces deux structures inhibitrices ont une activité tonique, c'est à dire qu'elles inhibent en permanence leurs cibles : certains noyaux du Thalamus qui à leur tour se projettent sur le cortex frontal. Ainsi, l'expression de l'action est par défaut inhibée ; c'est uniquement une activité de sélection dans les ganglions de la base qui va aboutir à la sélection d'une unité sur la couche de sortie. Cette unité va désinhiber sélectivement l'action correspondante dans le pôle moteur du cortex. Cette sélection est réalisée sur la base de l'information contextuelle perceptivo-motrice reçue en couche d'entrée et de la prédiction de la récompense faite par la seconde boucle des BG.

La boucle striato-nigrale (Haber et al., 2000) relie réciproquement le Striatum avec une structure appelée SNc (Substance Noire pars compacta) et correspond au critique. SNc est une des rares structures cérébrales composées de neurones dopaminergiques, la dopamine (DA) étant le neurotransmetteur modulateur lié à l'expression de la récompense. De façon schématique, on peut dire que SNc reçoit du Striatum (et d'autres structures cérébrales) des informations relatives à la récompense et va en retour influencer de façon modulatrice l'activité du striatum et ainsi la boucle basale. Depuis les travaux fondateurs de W. Schultz, il semble en effet clair que la dopamine code l'erreur de prédiction de la récompense (Schultz, 1997) et cela donne à ce mécanisme une proximité certaine avec l'algorithme des différences temporelles (Sutton, 1988).

Ce résumé fonctionnel souligne l'analogie entre les deux boucles constitutives des BG et l'architecture Acteur-Critique de l'Apprentissage par Renforcement. De nombreux travaux ont été menés pour préciser (ou relativiser) cette analogie. Concernant la boucle basale, la question principale est relative aux critères de sélection de l'action par désinhibition d'une unité de la couche de sortie à partir de l'information d'entrée. S'il y a effectivement une projection directe de la couche d'entrée, le Striatum, vers la couche de sortie, GPi/SNr, d'autres noyaux composent les BG (comme STN évoqué plus haut ou un noyau interne GPe) et permettent d'autres chemins que cette voie directe entre entrée et sortie, comme une voie indirecte (Gurney et al., 2001) ou une voie hyperdirecte (Leblois et al., 2006) et de nombreux travaux étudient des interactions possibles entre ces voies pouvant aboutir à une désinhibition sélective de l'action (Gurney et al., 2001). D'autres travaux étudient la représentation de l'information le long de cette boucle. D'une part, l'accent peut être mis sur la ségrégation de l'information selon des aspects relatifs à différents niveaux de la sélection de l'action (stratégie, planification, exécution correspondant à des

compétences de motivation, planification et action (Cisek, 2005)) représentés topologiquement et conservés sous forme de canaux différenciés le long de la boucle (Alexander et al., 1986). D'autre part, de façon quelque peu antagoniste, on peut mettre en avant la très petite taille de la couche de sortie GPi/SNr (10^5 neurones chez les primates, donc dix-mille fois plus petite que l'entrée corticale!) et parler plutôt d'un effet d'entonnoir et donc d'une importante réduction de la dimensionnalité entre l'entrée et la sortie (Bar-Gad et al., 2003).

Concernant la boucle striatonigrale (le critique), les travaux actuels se tournent principalement vers la meilleure compréhension de la proximité entre le comportement temporel de cette boucle et l'algorithme de différences temporelles, que l'on peut chercher à raffiner pour mieux correspondre à la version biologique (Daw et Doya, 2006).

2. Réduction de dimensionnalité

2.1. Dans les réseaux de neurones artificiels

De façon générale, la réduction de l'information est un mécanisme de filtrage bien connu en traitement automatique de l'information et peut se traduire par la réduction du nombre de données, par exemple par un mécanisme de clustering résumant un ensemble de données par un prototype représentatif, ou par la réduction de la dimensionnalité de l'information, comme c'est le cas dans l'Analyse en Composantes Principales (ACP). Ces deux mécanismes ont été décrits dans le domaine des réseaux de neurones artificiels. Concernant l'ACP (Diamantaras et Kung, 1996), il est connu depuis longtemps que si l'on applique la loi d'apprentissage neuronale classique, la loi de Hebb (Eq. 1), à la modification des poids entre une couche de neurones d'entrée X de dimension m et un unique neurone de sortie y (Eq. 2), cela va, au cours de l'apprentissage, extraire une direction s'alignant sur la première composante principale de l'espace d'entrée.

$$\Delta W_{ij} = \alpha(x_i y_j) \quad (1)$$

$$\Delta W = \alpha X y \quad (2)$$

où α est un petit réel positif, le pas d'incrémentation.

Cependant, cette loi est également connue pour diverger, ce qui rend l'accès à la direction extraite difficile. Une solution classique pour éviter cette divergence est de normaliser la loi, par exemple en divisant par la norme du poids, mais le calcul n'est plus local, ce qui peut être gênant dans un cadre neuromimétique. C'est pourquoi E. Oja a proposé de linéariser cette normalisation en utilisant, par approximation, le premier terme de la série de Taylor correspondante (Oja, 1982), ce qui a également l'avantage de rendre le calcul local (Eq. 3).

$$\Delta W = \alpha(Xy - Wy^2)$$

(3)

Cette règle d'apprentissage est donc stable et converge (si α est choisi suffisamment petit) vers un ensemble de poids correspondant à la direction de la première composante principale, pour un unique neurone de sortie. Des travaux ultérieurs ont étudié la possibilité d'extraire plusieurs composantes principales, en disposant plusieurs neurones dans la couche de sortie Y, dotée d'une connectivité latérale représentée par une matrice de poids A. Les neurones de la couche de sortie Y sont évalués linéairement par la somme pondérée des activités montantes et latérales (Eq. 4).

$$y_i = \sum_{j=1}^{m} W_{ij} x_j + \sum_{j=1}^{n} A_{ij} y_j \tag{4}$$

Le point commun de ces travaux est d'utiliser, entre ces neurones Y, un apprentissage anti-hebbien (Carlson, 1990 ; Zufiria et Berzal, 2007) dont le principe est de décorréler les activations de ces sorties (Eq. 5).

$$\Delta A_{ij} = -\alpha y_i y_j, i \neq j \tag{5}$$

α étant un petit réel positif et A la matrice des poids latéraux dans la couche de sortie Y.

Généralement, les composantes principales sont extraites successivement, par ajout incrémental de neurones dans la couche de sortie ou en définissant la matrice A comme une matrice triangulaire inférieure avec diagonale nulle, ce qui impose de fait une relation hiérarchique entre les neurones de sortie (Kung et Diamantaras, 1990 ; Rubner et Schulten, 1990 ; Carlson, 1990). Foldiak a également montré (Foldiak, 1989) qu'une couche de sortie complète proposée dès le début des calculs avec connectivité latérale complète permettait d'engendrer le sous-espace principal de dimension correspondante (sans avoir cependant accès aux directions principales individuelles). Précisons enfin que ces réseaux sont généralement dotés de neurones linéaires, afin de reproduire l'ACP, elle-aussi linéaire. Certains modèles cependant explorent des versions non linéaires du fonctionnement neuronal (Carlson, 1990), afin de réaliser des sortes d'ACP non-linéaires correspondant à des statistiques d'ordre supérieur (Diamantaras, 2002).

2.2. Dans les ganglions de la base

Aussi étrange que cela puisse paraître, la réduction de dimensionnalité impressionnante qui s'opère sur la voie basale directe entre le cortex et GPi/SNr a très peu été exploitée par les activités de modélisation. Seuls I. Bar-Gad et ses collègues (Bar-Gad et al., 2003) ont effectivement proposé qu'une forme d'ACP pourrait être réalisée entre ces couches. Un de leurs arguments importants est que les modèles classiques de sélection de l'action nécessitent une forte compétition latérale entre les neurones le long de la voie basale directe Cortex-Striatum-GPi/SNr, alors que des travaux d'électrophysiologie (Jaeger et al., 1994) reportent des poids latéraux très

faibles dans la partie basale de cette voie. En revanche, si on postule un traitement de type ACP, son action revient à décorréler les activités des neurones et, lorsque la fonction est apprise, la situation tend effectivement vers des poids latéraux nuls.

En proposant son modèle RDDR (Reinforcement Driven Dimensionality Reduction) de la voie basale directe, (Bar-Gad et al., 2003) reprend le modèle APEX d'ACP proposé par (Kung et Diamantaras, 1990), avec des poids montants régis par la loi de Oja et une couche de sortie hiérarchique avec des poids latéraux adaptés selon une loi anti-hebbienne, également adaptée de la règle de Oja.

L'originalité principale de ce modèle RDDR consiste à proposer que l'apprentissage du flux montant pourrait être modulé par le niveau de récompense reçu pour l'exemple considéré, exploitant ainsi le rôle modulateur de la voie dopaminergique portée par la boucle striato-nigrale, sur la boucle basale directe. Ainsi, les poids montants sont modifiés selon l'Eq. 6.

$$\Delta W_{ij} = \alpha r(x_i y_j - W_{ij} y_i^2) \tag{6}$$

où r est la récompense associée à l'exemple considéré, tandis que les poids latéraux sont mis à jour selon l'Eq. 7.

$$\Delta A_{ij} = -\alpha(y_i y_j + A_{ij} y_i^2), i > j, A_{ii} = 0 \tag{7}$$

La mise en œuvre et l'évaluation du modèle RDDR sont restées très formels et très sommaires et font partie des aspects que nous avons voulu améliorer dans le travail que nous rapportons ici. En effet, le modèle RDDR a uniquement été évalué sur sa capacité à effectuer une ACP conditionnellement à une récompense. Pour cela, des stimuli simples sont construits, correspondant à des matrices 8x8 où seules une ligne et/ou une colonne sont à 1, le reste de la matrice étant nul. Le but du réseau est d'apprendre à construire une représentation réduite des entrées, conditionnellement à la récompense. Pour cela, la récompense sera tout d'abord associée à la présence d'une colonne activée dans la matrice et dans une deuxième phase à la présence d'une ligne activée. Comme on le voit aisément, dans les deux cas, l'apprentissage de la transformation peut se faire sans perte d'information avec une sortie de taille 8 et c'est principalement cette compétence qui va être évaluée pour RDDR, de façon quelque peu artificielle. En effet, l'information obtenue en sortie va être décompressée en la reprojetant sur une couche artificielle de la même taille que la couche d'entrée, au moyen de la matrice inverse des poids calculés. En dehors de cette étude purement formelle (et possible uniquement car la transformation est linéaire), d'autres indices sont également mesurés lors de la convergence de l'apprentissage et selon le type de récompense donné. De façon très intéressante, il est en effet observé que la corrélation entre les neurones de la couche de sortie, ainsi que la valeur des poids latéraux dans cette même couche (matrice A), diminuent au cours de l'apprentissage, illustrant ainsi la décorrélation entre les directions extraites par les neurones

de sortie. Lorsque le critère de récompense change soudainement (la récompense passe des lignes aux colonnes), ces valeurs vont tout à coup augmenter rapidement pour décroître à nouveau lors du nouvel apprentissage. Le niveau de stabilisation de cet apprentissage dépend éventuellement du bruit introduit dans le fonctionnement (voir (Bar-Gad et al., 2003) pour des figures illustrant ces convergences successives). Nous soulignons ici que la redéfinition complète de la représentation dans GPi/SNr lorsque la règle de récompense change peut être comparée aux observations rapportées dans (Pasquereau, 2007), qui souligne l'extrême volatilité des réponses de ces neurones lors de protocoles de conditionnement.

2.3. Problématique du présent travail

Avec le modèle RDDR comme point de départ, nous avons voulu y apporter des modifications visant à complexifier ce mécanisme pour le rendre plus plausible biologiquement, c'est-à-dire plus proche de certaines caractéristiques fondamentales établies dans la biologie de la sélection de l'action, tout en désirant observer si cette propriété essentielle de réduction de dimensionnalité résistait à ces modifications. En résumé, elles portent sur les points suivants :

- Utilisation de champs de neurones dynamiques comme unités de calcul avec en particulier mécanisme de fuite et non-linéarité
- Ajout d'un axe cortical sensorimoteur permettant de préactiver les actions éligibles
- Fermeture de la boucle basale, avec retour vers le cortex moteur
- Définition d'un protocole d'apprentissage plus écologique
- Envoi de la récompense par l'environnement comme résultat de la sélection puis du déclenchement de l'action
- Ajout d'un mécanisme d'exploration

Nous précisons ces modifications dans la section suivante. Nous préciserons également en conclusion pourquoi nous pensons que ces travaux peuvent avoir un impact dans le domaine de l'apprentissage automatique.

3. Un modèle bio-inspiré d'apprentissage par renforcement

Les extensions apportées au modèle RDDR sont les suivantes :

3.1. Champs neuronaux dynamiques

Le modèle RDDR utilise des modèles de neurones formels, évaluant à chaque cycle le nouvel état de chaque unité. Nous avons choisi d'utiliser le formalisme classique des modèles bio-inspirés, les champs de neurones

dynamiques (Amari, 1977 ; Erlhagen et Schoener, 2002), où l'état d'activation *u* des unités est contrôlé par une équation différentielle (cf. Eq. 8 pour sa version discrétisée effectivement employée pour les simulations) avec un terme de fuite, une non-linéarité représentée par la fonction f et le paramètre $0<\delta<1$ permettant d'assurer que la dynamique est contractante. Dans cette même équation, k est l'indice permettant d'itérer à l'intérieur du champ neuronal (c'est-à-dire représentant la connectivité latérale) et j est un indice sur les structures en entrée de ce champ (ici sur le flux montant). h représente un terme d'activité de base ou de bruit que nous utiliserons par la suite.

$$u_i(t+1) = f\left[u_i(t) + \delta\left[-u_i(t) + \sum_j W_{ij}x_j + \sum_k A_{ik}u_k + h\right]\right] \quad (8)$$

3.2. Axe cortical sensorimoteur

Nous avons souhaité prendre en compte le fait que la sélection de l'action ne se fait pas sur l'ensemble des actions potentielles que nous pouvons faire mais sur un sous-ensemble restreint d'actions suggéré par la scène perceptive et préactivé dans le cortex moteur par l'axe associatif cortical pariétal (Goodale et Humphrey, 1998), répondant ainsi à la notion d'affordance (Gibson, 1979). Ainsi, au cours de l'apprentissage par renforcement effectué dans la boucle basale, un apprentissage associatif va simplement stocker pour chaque perception, dans une aire associative correspondant au cortex pariétal, l'ensemble des actions qui se sont vues dans le passé associées à cette perception. Ce sous-ensemble d'actions potentielles sera, au début de la sélection de l'action, maintenu sous le seuil de déclanchement par l'activité inhibitrice tonique (par défaut) de la couche de sortie des BG, mais cette préactivation sera suffisante pour activer le Striatum. La tâche principale de la boucle basale sera de sélectionner et d'activer pleinement une action à déclencher effectivement et d'inhiber les autres. Soulignons que l'ajout de ce mécanisme est uniquement motivé par nos discussions avec nos collègues physiologistes, en particulier pour comparer les décours temporels des structures neuronales impliquées et qu'a priori il ne change rien sur le principe de réduction de dimensionnalité. Au contraire, il rendra plus difficile le changement de politique de sélection et nécessitera l'introduction d'un mécanisme d'exploration (cf. 3.6).

3.3. Fermeture de la boucle basale

La principale modification que nous avons apportée au modèle RDDR original a été de réaliser la boucle basale complète (Cortex-Striatum-GPi/SNr-Thalamus-Cortex) afin de voir si le processus de réduction de dimensionnalité continuait d'émerger de cette boucle fermée, plus naturelle et plus dynamique. En particulier, comme nous utilisons maintenant le formalisme des champs neuronaux dynamiques qui évalue les unités

itérativement en prenant en compte leur état précédent, il était de première importance d'observer comment la sélection de l'action émergeait progressivement, partant du stimulus initial et des actions préactivées correspondantes, par la réentrance via le thalamus de la dynamique de la boucle.

Il est intéressant de souligner que, dans cette nouvelle configuration, l'aire motrice est maintenant au centre de ce réseau plus complet que nous avons construit (cf. Fig. 1), au croisement de la boucle basale et de l'axe cortical sensorimoteur. Chaque unité de l'aire motrice met à jour son activité à partir de son terme de fuite, de l'entrée corticale associative et du feed-back de la boucle basale, transmis par GPi/SNr via le thalamus. Cette activité mise à jour est ensuite transmise (ainsi que l'activité sensorielle) au Striatum et se projette ainsi sur les composantes principales courantes, jusqu'à ce qu'une unité de l'aire motrice franchisse un seuil et déclenche son action en inhibant les autres.

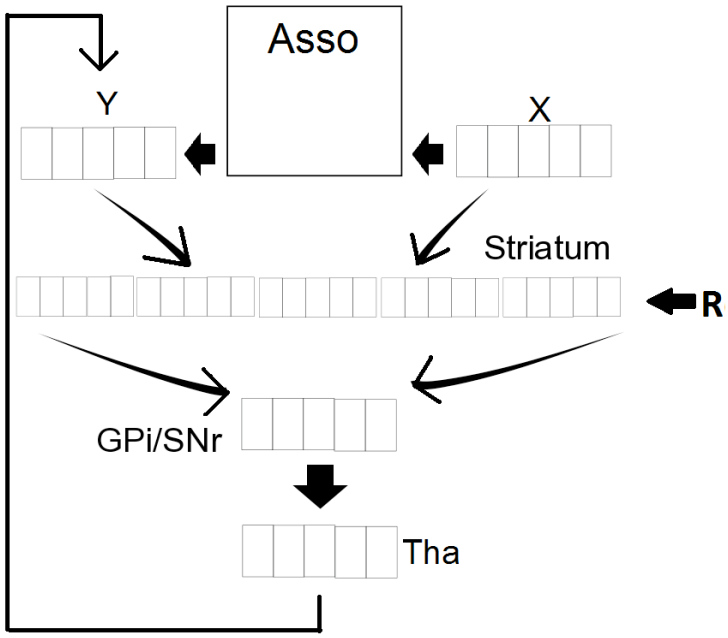

Fig. 1 : Architecture générale du réseau. X : cortex sensoriel ; Asso : cortex associatif ; Y : cortex moteur ; Tha : noyau ventro-médian du thalamus ; R : récompense (venant de SNc).

Conformément à des données biologiques que nous ne décrirons pas ici, les unités du Striatum, du Thalamus et du cortex moteur sont non-linéaires et une inhibition latérale fixe est imposée dans l'aire motrice.

3.4. Protocole d'apprentissage

Le protocole d'apprentissage que nous avons mis au point n'est pas plus complexe que celui utilisé dans (Bar-Gad et al., 2003), mais il est plus écologique (plus naturel) du point de vue du conditionnement opérant. Un parmi cinq stimuli différents est proposé à un sujet, qui doit répondre en déclenchant une action parmi cinq possibles (par exemple bouger son bras pour atteindre un bouton). Une loi de récompense (cachée au sujet) associe une action à chaque stimulus. Deux lois ont été implantées ; la première associe le stimulus i à l'action i ; la seconde échange les actions 2 et 3, ainsi que les actions 4 et 5. Le changement de loi de récompense se fait sans avertissement au sujet.

3.5. Dynamique temporelle de la récompense

Dans les expérimentations réalisées par (Bar-Gad et al., 2003), une récompense était systématiquement associée à chaque stimulus (les lignes ou les colonnes de la matrice). Ici, afin de définir un protocole plus écologique, la récompense est donnée seulement quand l'action a été déclenchée (et sous réserve qu'elle corresponde à la loi de récompense), comme résultat de son impact supposé dans l'environnement. En résumé, un stimulus est initialement proposé et préactive un certain nombre d'actions potentielles. Ces deux informations sont envoyées dans la boucle basale qui va boucler sur elle-même jusqu'à ce qu'une action soit sélectionnée. Selon la loi de récompense courante, la valeur de la récompense sera 0 ou 1 et la matrice de poids W sera mise à jour en conséquence. Si une nouvelle loi est appliquée, une nouvelle action pourra être choisie, même sans préactivation, grâce au mécanisme d'exploration que nous décrivons maintenant.

3.6. Mécanisme d'exploration

L'équation des champs neuronaux (Eq. 8) inclut un terme (h) pouvant correspondre au bruit ou à l'activité spontanée. De telles fluctuations sont une caractéristique importante des systèmes neuronaux et permettent d'obtenir des réponses non déterministes. Ceci est particulièrement intéressant en conditionnement opérant où un mécanisme d'exploration est très souvent utile. De plus, dans notre étude, le mécanisme de préactivation doit être compensé par la possibilité de choisir des associations inédites. Nous avons implanté un tel mécanisme dans la couche représentant GPi/SNr, pour le choix critique de l'action à désinhiber : une variable aléatoire centrée et suivant une loi normale est ajoutée lors de l'évaluation de chaque unité, avant de choisir l'unité la plus active, lors de l'évaluation non-linéaire du thalamus qui va à son tour contribuer à l'activité du cortex moteur.

4. Simulations et premiers résultats

Nous présentons ici les premiers résultats obtenus avec cette architecture plus complète et un protocole d'apprentissage plus réaliste.

4.1. Structure du réseau

Les structures cérébrales impliquées dans la tâche considérée ont été modélisées à l'aide de matrices et de vecteurs. Plus précisément (cf. Fig. 1), le cortex moteur Y est représenté par un vecteur de réels de dimension 5 ; le cortex sensoriel X par un vecteur de dimension 5, à valeurs binaires (1 quand la perception est présente ; 0 sinon) ; le cortex pariétal associatif par la matrice binaire Asso de taille 5*5 ; le Striatum par un vecteur de réels de dimension 25 ; GPi/SNr par un vecteur de réels de dimension 5 et le noyau ventro-médian du thalamus par un vecteur binaire de dimension 5. Les non-linéarités dans le cortex moteur et le striatum sont obtenues par la fonction *tanh*.

4.2. Comportement du réseau

La tâche du réseau est de répliquer le protocole décrit dans la section 3.4 plus haut, où la loi de récompense peut être modifiée sans avertissement et où le sujet doit alors trouver seul la nouvelle loi. Sur cette base, de nombreux exemples d'apprentissage sont fournis au réseau et on observe comment le réseau apprend à associer perceptions et actions, conditionnellement à la récompense.

En résumé, nous avons effectivement observé qu'après quelques milliers d'exemples, la relation était apprise et le comportement correct généré, sauf quand le mécanisme d'exploration choisissait une autre action. Afin de mesurer plus précisément le comportement du réseau, nous avons pour le moment réalisé deux sortes d'évaluation.

	W_1	W_2	W_3	W_4	W_5
W_1*	**13.8**	**76.3**	**91.3**	**90.1**	**91.0**
W_2*	**90.7**	**26.4**	**65.1**	**82.2**	**91.3**
W_3*	**90.0**	**89.9**	**15.8**	**104.3**	**96.7**
W_4*	**90.1**	**90.3**	**90.3**	**1.4**	**91.3**
W_5*	**89.9**	**90.7**	**90.1**	**89.8**	**0.77**

Table 1 : Angles (en degrés) entre les directions calculées (colonnes) et théoriques (lignes), pour la première loi de récompense.

Premièrement, comme reporté dans la Table 1, nous avons mesuré les angles entre les directions principales calculées et désirées. A cause des non-linéarités introduites et du fait que la boucle est maintenant fermée, il

n'est plus possible de réaliser l'opération (artificielle) consistant à reconstruire la représentation originale en décompressant la représentation apprise. En revanche, la simplicité de l'association à apprendre nous permet de calculer analytiquement les directions principales théoriques et de les comparer aux directions extraites dans les vecteurs de poids des lignes de la matrice W. Ce sont les vecteurs créés par les poids liant la couche du Striatum à chaque neurone de GPi/SNr et ils sont supposés extraire les composantes principales et être orthogonaux entre eux. Pour la première loi de récompense, nous avons comparé (Table 1) ces directions apprises et théoriques. Nous observons que les angles entre ces vecteurs désirés et appris sont effectivement faibles et que les angles entre les directions principales sont effectivement proches de 90°. Nous pensons que les erreurs commises sont principalement dues au mécanisme d'exploration.

Deuxièmement, il était aussi important de vérifier que la matrice de poids latéraux inhibiteurs A converge vers 0 au cours de l'apprentissage. Ces connexions inhibitrices sont responsables de l'orthogonalité des directions principales et réciproquement leur valeur nulle indique que l'orthogonalité est réalisée dans W. Comme représenté dans la Fig. 2, cette convergence se réalise bien au cours de l'apprentissage, pour une même loi de récompense, mais ne va pas exactement à 0 à cause du mécanisme d'exploration.

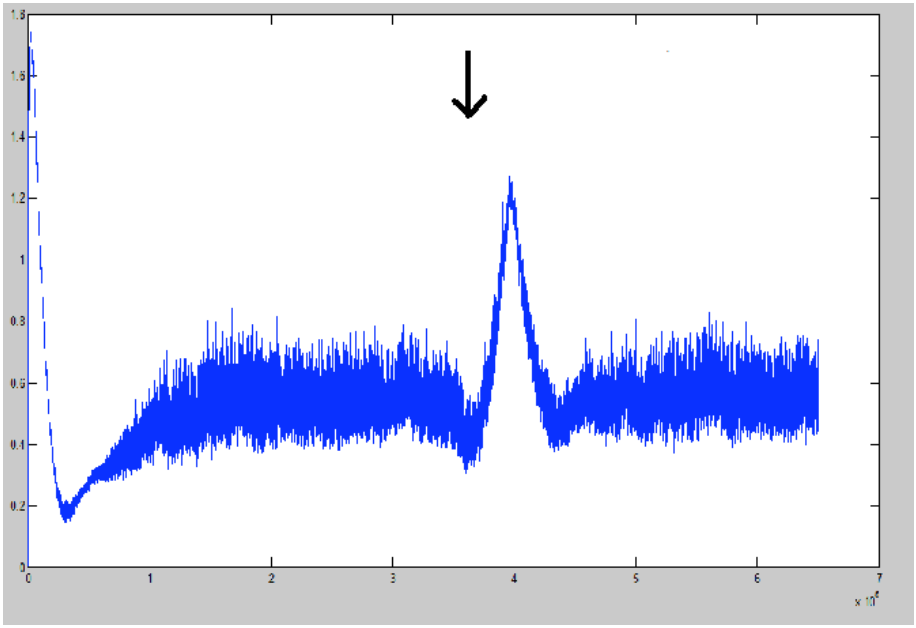

Fig. 2 : Evolution (en ordonnée) de la norme de la matrice A (poids latéraux inhibiteurs) au cours de l'apprentissage (en abscisse, nombre d'exemples présentés). Le changement de la loi de récompense (indiqué par la flèche) entraine une augmentation de la norme, qui décroît ensuite à nouveau au cours de l'apprentissage. La norme n'atteint pas 0 à cause du mécanisme d'exploration.

Il est encore plus intéressant d'observer (toujours dans la Fig. 2) l'évolution de la norme de la matrice A quand la loi de récompense est modifiée et de constater cette augmentation brutale de la norme qui va à nouveau décroître quand l'apprentissage sera poursuivi et les nouvelles directions principales extraites. Quand une association perception-action inédite est faite, le mécanisme d'exploration va permettre d'activer d'autres actions, d'en découvrir la récompense associée et de l'apprendre.

5. Conclusion

5.1. Approche et principaux résultats

L'apprentissage par renforcement reste aujourd'hui un mécanisme adaptatif mal maîtrisé, en particulier à cause de la combinatoire sous-tendue par la plupart des algorithmes existants et à cause d'une prise en compte trop sommaire de certains de ses composants essentiels, comme la récompense. L'inspiration du vivant est une voie possible pour mieux observer, au niveau comportemental, les caractéristiques de cet apprentissage et, au niveau physiologique, le substrat neuronal qui le réalise. Ceci peut sembler d'autant plus pertinent que des travaux préalables ont commencé à bien identifier les circuits correspondants et à montrer en particulier l'implication de la boucle basale.

Le modèle RDDR a lancé une nouvelle direction de recherche, en postulant une réduction de dimensionnalité de type ACP dans la boucle basale. Une règle d'apprentissage originale associée, modulant l'extraction des composantes principales par un signal de renforcement, a également été présentée dans (Bar-Gad et al., 2003). Dans le présent papier, nous proposons d'étendre ce modèle à un formalisme plus réaliste de calcul neuronal, à une structure de réseau plus réaliste et à un protocole d'apprentissage plus réaliste. En dépit de ces modifications majeures, nous montrons ici que ce mécanisme de réduction de dimensionnalité modulée par le renforcement fonctionne toujours efficacement et que la sélection de l'action est toujours de bonne qualité.

D'une part, nous pensons qu'un tel système plus complet est un meilleur substrat pour collaborer avec des neurobiologistes à la compréhension fine de la dynamique de ce réseau et de l'évolution de ses représentations, ce que nous commençons actuellement. D'autre part, les mécanismes calculatoires présentés ici pourraient également avoir un impact dans le domaine de l'apprentissage automatique. De façon ciblée, la réduction de la dimensionnalité ou la présélection des actions ont par exemple un effet sur la combinatoire des associations explorées. De façon plus générale, fermer la boucle du comportement, souligner le lien fort entre conditionnement opérant et répondant ou mieux définir le codage de la récompense

pourraient être des sujets de réflexion commune entre ces différentes approches de modélisation de l'apprentissage par renforcement.

5.2. Perspectives

D'autres travaux en cours consistent à réduire les problèmes dus au mécanisme d'exploration, en rendant cette exploration dépendante du nombre cumulé d'essais sans récompense. Ceci pourrait être expliqué par une tendance à essayer aléatoirement des réponses si peu de récompense a été obtenu depuis un certain nombre d'essais. Inversement, si des récompenses sont reçues régulièrement, cette tendance à l'exploration pourrait décroître.

Sur un plus long terme, nos perspectives de travail sont de deux ordres, toutes deux orientées vers un nouvel accroissement de la plausibilité biologique de notre système pour mieux comprendre les ingrédients de l'apprentissage par renforcement. D'une part, nous pensons que la comparaison avec la biologie sera plus profonde si la taille du réseau est plus importante. En particulier, des couches comportant plus d'unités pourraient développer des phénomènes d'auto-organisation topologique, comme il est observé dans la plupart des structures concernées ici (Haber et al., 2000).

D'autre part, on peut remarquer que la plupart des travaux rapportés ici sont relatifs à la partie Acteur de l'architecture. Développer un Critique plus réaliste est aussi de première importance pour améliorer le système. Les questions principales à explorer concernent l'approfondissement du rôle de la dopamine dans ce système et l'ajout à ce système de conditionnement opérant d'un composant de conditionnement répondant plus réaliste (O'Reilly et al., 2007). Enfin, nous espérons qu'une compréhension plus fine de ces mécanismes permettra de générer un débat utile avec la communauté d'apprentissage automatique. Elle nous apporte un cadre utile pour l'analyse de nos modèles ; nous espérons en retour lui apporter une vision originale et riche en inspiration.

Références

Alexander, GE, DeLong, MR, Strick, PL (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex, *Ann. Rev. Neurosci,* 9:357-81.

Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields, *Biological Cybernetics* 27 (2) (1977) 77–87.

Bar-Gad I., Havazelet Heimer G., Goldberg J. A., Ruppin E., and Bergman H. (2000) Reinforcement driven dimensionality reduction - a model for information processing in the basal ganglia. *Journal of Basic & Clinical Physiology & Pharmacology* 11:4.

Bar-Gad, I, Morris, G., Bergman, H. (2003) Information processing, dimensionality, reduction and reinforcement in the basal ganglia. *Progr. Neurobiol.*71:439-477

Bar-Gad, I, Bergman, H. (2006) Reinforcement driven nonlinear dimensionality reduction in the multilayer network of the basal ganglia. In: *Recent Breakthroughs in Basal Ganglia Research*, Erwan Bezard ed, pp 45-52, Nova Science Publishers.

Carlson, A. (1990). Anti-Hebbian learning in a nonlinear neural network, *Biol. Cybern*., vol. 64, pp. 171–176.

Cisek, P. (2005) "Neural representations of motor plans, desired trajectories, and controlled objects". *Cognitive Processing*. 6: 15-24.

Craig, A.D., How do you feel - now? The anterior insula and human awareness, *Nat. Rev. Neurosci*. 10, pp. 59-70, 2009.

Daw, ND and Doya, K (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16:199–204.

Diamantaras, K. I. & Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications*. Toronto: Wiley.

Diamantaras, K.I. Neural Networks and Principal Component Analysis, Chapter 8 of *Handbook of Neural Network Signal Processing*. Yu Hen Hu and Jeng-Neng Hwang, Editors. CRC Press, New York, 2002.

Erlhagen, W. and Schoener, G. Dynamic field theory of movement preparation, *Psychol Rev* 109 (3) (2002) 545–72.

Fix, J. A computational approach to the control of voluntary saccadic eye movements. *International Conference on Cognitive Neurodynamics,* ICCN-2007. (2007).

Fix, J. and Rougier, N. and Alexandre, F., From physiological principles to computational models of the cortex, *Journal of Physiology*, 101, 1-3, pp. 32-39, 2007.

Földiák, P. (1989) Adaptive network for optimal linear feature extraction, *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, Washington DC., vol. 1, pp. 401-405 (IEEE Press, New York).

Frezza-Buet, H. and Alexandre, F., From a biological to a computational model for the autonomous behavior of an animat, *Information Sciences*, 144(1-4), p. 1-43, Jul. 2002.

Gibson, J.J. (1979). The Ecological Approach to Visual Perception. Boston : Houghton Mifflin.

Goodale, M.A., Humphrey, G.K.: "The objects of action and perception", *Cognition* 67, pp. 181-207, 1998.

Gurney, K., Prescott, T.J., and Redgrave, P. (2001). "A computational model of action selection in the basal ganglia. I. A new functional anatomy." *Biol. Cybern*. 84: 401-410.

Haber, SN., Fudge, JL and McFarland, NR (2000). Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum; *The Journal of Neuroscience*, 20(6):2369–2382.

Jaeger, D., Kita, H., Wilson, C.J., 1994. Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum., *Journal of Neurophysiology*, 72, 2555–2558.

Joel, D., Niv Y. and Ruppin, E. (2002) - Actor-critic models of the basal ganglia: New anatomical and computational perspectives - *Neural Networks* 15, 535-547.

Kohonen, T., *Self-Organizing Maps*, New York : Springer-Verlag, 1997.

Kung, S.Y., Diamantaras, K.I., (1990). A neural network learning algorithm for adaptive principal component extraction (APEX). *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.* 2, 861--864.

Leblois, A., Boraud, T., Meissner, W., Bergman, H., and Hansel, D. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J. Neurosci.* 26, 3567-3583.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C.E., and Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *J Neurosci* 27, 1176-1183.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J.Math.Biol.,* 15, 267-273.

Redgrave, P., Prescott, T.J. & Gurney, K. (1999), The basal ganglia: a vertebrate solution to the selection problem?, *Neuroscience*, 89:1009-1023.

Rubner J, Schulten K (1990) Development of feature detectors by self-organization: A network model. *Biol Cybern* 62:193-199.

Schultz W, Dayan, P, Montague RR. A neural substrate of prediction and reward. *Science* 275: 1593-1599, 1997

Sutton, R.S., 1988, Learning to Predict by the Method of Temporal Differences, *Machine Learning*, 3, pp. 9-44.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press Cambridge, MA.

Zufiria, PJ and Berzal, JA (2007) Analysis of Hebbian Models with Lateral Weight Connections ; F. Sandoval et al. (Eds.): *International Workshop on Artificial Neural Networks, IWANN 2007*, LNCS 4507, pp. 39–46, Springer-Verlag.